CODATA, IUCr, PDBj collaboration for medical-protein crystal structure definitive versions of data files

Helliwell, J R*, Kurisu, G, and Kroon-Batenburg, L

University of Manchester, UK; University of Osaka and PDBj, Japan; University of Utrecht, The Netherllands

Introduction

At the Research Data Alliance Plenary 17 (https://www.rd-alliance.org/plenaries/rda-17th-plenary-meeting-edinburgh-virtual) concerns were voiced that "multiple versions of covid-19-proteins crystal structure data were useless". I (JRH) replied that it was a form of mayhem but not useless. But how to improve? The multiple versions had arisen from well-meaning multiple task forces offering improved versions on their own personal websites but in the end largely ignored by the depositors at the PDB with only 30 revised versions out of more than 1000 depositions as of IUCr Prague Congress. Clearly this suggested to JRH that a direct collaboration with the PDB would be an improved approach.

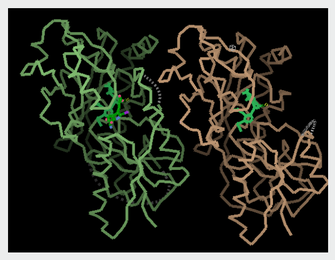
Aim

So, envisaging a process akin to a journal's peer review we set up a collaboration within the CODATA GOSC Case Studies, formally endorsed by the IUCr (see https://codata.org/initiatives/decadal-programme2/global-open-science-cloud/case-studies/diffraction-data/). Progress of this initiative has been made and spans covid-19 and other medically important proteins (*e.g.* see Helliwell 2021).

Method

That PDBj had launched a raw diffraction images data archive XRDa https://xrda.pdbj.org/ was pivotal as it would allow a combined evaluation of raw data, processed structure factors and derived protein molecular model. This also would lead to general community benefit beyond medical pandemic challenges, although of course very important, to the whole of macromolecular crystallography. Feedback on a PDBj deposition is made by JRH and LKB to GK and who then can decide, like a journal editor exactly what feedback is made to a depositor to PDBj for a possible reversioning of a PDBj deposition.

Reprocessing results on a test case, 7ccy



This is an exemplary case:-

No difference map peaks at 5.00 sigma

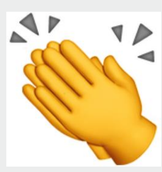


Table 1 Let's check the resolution limit various indicators

Dials reprocessing in ccp4i2 (Beilsten-Edmands et al 2020) using the PDBj 7ccy diffraction images as an example (Sato et al 2021). It gives a useful range of diagnostics of the possible resolution limits (and what a large range those possible resolution limits are!). Resolution cut off estimates:-

resolution of all data	: 1.913
based on CC(1/2) >= 0.33	: 1.946
based on mean(I/sigma) >=2.0	: 3.037
based on R-merge < 0.5	: 2.411
based on R-meas < 0.5	: 2.497
based on completenes >=90%	: 2.335
based on completeness >=50%	: 2.155

Using iMosflm reprocessed raw data to 2.0Å (Battye et al 2020) PDB Redo estimate based on ***Diederichs and Karplus***:-

******** Testing resolution cut-offs: 2.40*Å* 2.29*Å* 2.16*Å* 2.00*Å.***

o Testing resolution 2.40

o Testing resolution 2.29

*** LL-free deteriorated**

o Testing resolution 2.16

*** R-free deteriorated**

*** Weighted R-free deteriorated**

-High resolution cut-off:2.29Å

Conclusions

Dials makes several recommendations about the resolution limit for protein model refinement. Sato et al used a diffraction resolution limit of 2.40Å, which is a good choice across the several parameters listed above.

For clarity we suggest that it would be better to apply the best model as the criterion which Diederichs and Karplus have shown should be via the paired refinement method (Maly et al 2020 and references therein). This is available for instance at https://pdb-redo.eu/ (Joosten et al 2014).

Also, the (Fo-Fc) peaks' list should be acted upon before deposition in the PDB. To try to ensure this the PDB Validation Report could include a list of (Fo-Fc) peaks that have not been dealt with in the model to openly advise the depositor. That would ensure that the model likely does not need post publication peer review.

Acknowledgement

We are very grateful to Sato et al for depositing their raw, processed and derived data files at PDBj and XRDa.

References

Battye, T.G.G., Kontogiannis, L., Johnson, O., Powell, H.R. & Leslie, A.G.W. (2011) Acta Cryst. **D67**, 271-281.

Beilsten-Edmands, J., Winter, G., Gildea, R., Parkhurst, J., Waterman, D. & Evans, G. (2020). Acta Cryst. D76, 385-399.

Helliwell, J R (2021) Acta Cryst F77, 388-398.

Joosten RP, Long F, Murshudov GN, Perrakis A. IUCrJ. 2014 May 30;1(Pt 4):213-20.

Maly, M., Diederichs, K., Dohnalek, J. & Kolenko, P. (2020). IUCrJ 7, 681-692.

Sato, H., Sugishima, M., Tsukaguchi, M., Masuko, T., Iijima, M., Takano, M., Omata, Y., Hirabayashi, K., Wada, K., Hisaeda, Y. & Yamamoto, K. (2021). Biochem. J. 478, 1023–1042.